Kaouther HARBI, M2-Bioinformatics DLAD

Course: Biostatistiques approfondies

Cancer type prediction using Machine Learning algorithms

1. Introduction:

The continuous and rapid emergence of new types of cancer that affect more and more people with a high mortality rate is generating a particular interest of the scientific community in oncology.

An earlier diagnosis would be advantageous for the management of evolution of the pathology and crucial in the choice of the treatment methodology to be applied

Key elements in reduction of mortality rate among cancer carriers are: early detection, accurate determination of cancer histological type and adequate treatment. Errors in cancer type or, in general, malignant growth type determination led to treatment efficiency degradation, because anticancer strategy depends on tumor morphology (morphogenesis).

Traditionally, this diagnosis is performed manually by doctors, following biomedical analyses which consist of the determination of the cancer type/class and also its advancement stage.

As a goal to make the diagnosis more rapid and accurate, it is essential to reduce the core role played by the human and introduce more artificial intelligence. More precisely, machine learning (ML) for phenotyping approaches. This type of use of ML was first introduced around 2007[1].

Machine Learning is an technique of artificial intelligence which consists in "teaching" a machine, from data to make predictions.

In our case, ML models play a key role in taking into consideration effective features which conduct to a precise cancer type. In this review, we are interested into the phenotype determination (=cancer type) using gene expression levels of a variety of genes of interest) in the most occurring cancer types.[2]

2. Material and Methods:

2.1. Dataset:

The dataset is composed of the expression of 20531 genes of 801 patients with four different tumors. These four different tumors are breast cancer (BRCA), Colon Adenocarcinoma (COAD), Kidney Renal Clear Cell Carcinoma (KIRC), Lung Adenocarcinoma (LUAD) and Prostate Adenocarcinoma (PRAD). The data was obtained from the UCI Machine Learning Repository which contains 622 datasets for the Machine Learning community to practice. The goal here is to implement a model allowing the diagnosis of a cancer based on this dataset.[3]

2.2. Programming_language_

Python 3.9.0

2.3. Libraires:

- **Tensorflow** 2.7.0
- **Scikit-learn** 1.0.2
- **Scipy** 1.7.3
- **Pandas** 1.3.5
- **Seaborn** 0.11.2

2.4. Workflow:

The overall goal of this work was to predict phenotypes (output) based on the gene-expression(features): its then a classification challenge.

Different ML models were implemented, models which follow various strategies when it comes to classification of cancer-type.
First time, in each model implementation is data splicing into training and testing datasets.

X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.33, random_state=42, stratify=Y)

Then 5 classifications models were chosen based on a diversity criterion. We have here supervised learning algorithms based on a defined distance (KNN) or maximum likelihood (NB)or predicts a multinomial probability (MLR)or support vectors (SVM) or finally tree-based algorithms (RF).

- K-Nearest Neighbour (KNN)
- Naive Bayes (NB)
- Multiclass logistic regression (MLR)
- Support vector machines (SVM)
- Random Forest (RF)

A function (evaluationFunc) was established to compare the “quality” of the models which takes 2 parameters and with the use of the:

- Accuracy score
- Crosstab
- Visualization as a heatmap of the crosstab
- Detailed classification report

2.5. Optimization:

Different strategies were used to in order to optimize the models depending on the specificities of the models' parameters.

In K-Nearest Neighbour (KNN) one of the key parameters is the number of neighbors, aiming to determine it a grid search was used combined with a cross validation (method or resampling that uses different portions of the data to test and train a model on different iterations). The output of the GridSearch gave the number of neighbors which generates the best accuracy score.

In Multi class logistic regression a range of regularization terms C in [1e-10, 1e-5, 1, 1e5, 1e10, 1e20] were applied and their corresponding accuracy score was measured

3. Results and Discussion:

Most of the tested models had an “excellent” accuracy score (near to 1) when applied on the testing dataset. This could be caused by an overfitting issue that should be resolved by regularization penalty or a reduction in the number of the features taken into account. Another hypothesis is that the raw data used in this project is simply “too perfect” that the algorithm find ease in learning its specificities.

K-Nearest Neighbour (KNN) best parameters search were determined to be 1 neighbor taken into account and generating an accuracy score of 0.9962.

Naive Bayes (NB) came out to be the less accurate among the tested models with an accuracy score of 0.766. This is quite expected since this algorithm supposes the independence between the features, which is not the case since we are dealing with genes involved in cancers.

Fig1. Crosstab visualized as heatmap between the tested and the predicted values

The testing of different regularization terms in the **Multiclass logistic regression** led to determination of the value of C which leads to the best accuracy score which is around 1.

C	1e-10	1e-05	1	1e5	1e10	1e20
Accuracy score	0.37	0.98	0.99			

Support vector machines (SVM) showed one of the highest accuracy scores of 0.988 which can be caused to the kernel trick used by this type of classifiers when facing higher dimensions.

Random Forest (RF) also had a high accuracy score of 0.977 this could be due to the unsensitivity of such models to unnormalized data as it uses a rule-based approach.

A test of under-fitted model using less feature caused the accuracy score to decrease drastically (0.192).

4. Conclusion:

The different ML models in order to predict the corresponding phenotype according to the gene expression was made possible and this with a high accuracy score. But this should not be sufficient when it comes to measuring the effectiveness of the models since over-fitting problems were observed and even using regularization penalty, we still observed accuracy score near 1 for both training and testing datasets. Cross-validation and different regularization methods were used in order to see their corresponding effects. An overview of the 5 different models showed that the majority seemed to perform “well” while the Naive Bayes struggled achieving such accuracy.

The phenotype determination using ML algorithm is the future of automated diagnosis and could lead to future invention when combined to the machines performing the analysis especially with the availability of such enormous data quantities giving a faster answer with lower error risk since it takes into account an important number of features.

It could be valuable to test more robust and complex models in order to perform the same work and compare the accuracy results since it should interesting results in other studies[4].

5. Bibliography

1. Huang, Y., McCullagh, P., Black, N. & Harper, R. Feature selection and classification model construction on type 2 diabetic patients’ data. *Artif. Intell. Med.* **41**, 251–262 (2007).
2. SEER Cancer Statistics Review, 1975-2018. *SEER* https://seer.cancer.gov/csr/1975_2018/index.html.
3. UCI Machine Learning Repository: gene expression cancer RNA-Seq Data Set. https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

4. Grinberg, N. F., Orhobor, O. I. & King, R. D. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Mach. Learn.* **109**, 251–277 (2020).